2188

PTO/SB/21 (08-03)

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	09/825,607
Filing Date	April 3, 2001
First Named Inventor	Yamamoto, Yasutomo
Art Unit	2188
Examiner Name	Jasmine Song
Total Number of Pages in This Submission	
Attorney Docket Number	16869P-025300US

ENCLOSURES *(Check all that apply)*

- ☐ Fee Transmittal Form
  - ☐ Fee Attached
- ☐ Amendment/Reply
  - ☐ After Final
  - ☐ Affidavits/declaration(s)
- ☐ Extension of Time Request
- ☐ Express Abandonment Request
- ☐ Information Disclosure Statement
- ☒ Certified Copy of Priority Document(s)
- ☐ Response to Missing Parts/ Incomplete Application
  - ☐ Response to Missing Parts under 37 CFR 1.52 or 1.53
- ☐ Drawing(s)
- ☐ Licensing-related Papers
- ☐ Petition
- ☐ Petition to Convert to a Provisional Application
- ☐ Power of Attorney, Revocation Change of Correspondence Address
- ☐ Terminal Disclaimer
- ☐ Request for Refund
- ☐ CD, Number of CD(s)
- ☐ After Allowance Communication to Group
- ☐ Appeal Communication to Board of Appeals and Interferences
- ☐ Appeal Communication to Group (Appeal Notice, Brief, Reply Brief)
- ☐ Proprietary Information
- ☐ Status Letter
- ☒ Other Enclosure(s) *(please identify below)*:

  Return Postcard

Remarks

The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual	Townsend and Townsend and Crew LLP
	Steve Y. Cho — Reg. No. 44,612
Signature	
Date	3/17/04

CERTIFICATE OF TRANSMISSION/MAILING

I hereby certify that this correspondence is being facsimile transmitted to the USPTO or deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.

Typed or printed name	Andrea S. Beck		
Signature		Date	3/17/04

60167938 v1

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００１年　２月２８日

出願番号
Application Number: 特願２００１－０５３４７２

[ST. 10/C]: [ＪＰ２００１－０５３４７２]

出願人
Applicant(s): 株式会社日立製作所

２００４年　２月１９日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００４－３０１１０４９

【書類名】　特許願

【整理番号】　K00011751A

【提出日】　平成13年 2月28日

【あて先】　特許庁長官

【国際特許分類】　G06F 17/60

【発明者】

【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内

【氏名】　山本　康友

【発明者】

【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内

【氏名】　大枝　　高

【発明者】

【住所又は居所】　神奈川県小田原市国府津２８８０番地　株式会社日立製作所　ストレージシステム事業部内

【氏名】　佐藤　孝夫

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100075096

【弁理士】

【氏名又は名称】　作田　康夫

【手数料の表示】

【予納台帳番号】　013088

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報処理システム

【特許請求の範囲】

【請求項１】

ホストコンピュータと、前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置とを有する情報処理システムであって、

前記ホストコンピュータは、

前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録される情報記憶手段と、

前記複数のディスク装置のうち第一のディスク装置に記録されたデータを第二のディスク装置に移動する場合に、移動先の論理的な記憶領域及び前記情報記憶手段から、前記第二のディスク装置の物理的な記憶領域を示す情報を取得する取得手段と、

前記取得手段によって取得された前記第二のディスク装置の物理的な記憶領域を示す情報及び移動対象となるデータが格納されている前記第一のディスク装置の物理的な記憶領域を示す情報を、前記記憶装置に転送する転送手段とを有し、

前記記憶装置は、

前記転送手段によって転送された前記情報のうち、前記第一のディスク装置の物理的な記憶領域を示す情報を使用して、移動元のデータを読み出す手段と、

前記記憶手段に記憶された前記情報のうち、前記第二のディスク装置の物理的な記憶領域を示す情報を使用して、移動先のディスク装置に複写する複写手段と、

を有することを特徴とする情報処理システム。

【請求項２】

前記記憶装置は、

前記複写が終了したことを前記ホストコンピュータに通知する通知手段を有し、

前記ホストコンピュータは、

前記通知を受け取った後、前記情報記憶手段に登録されたディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係を更新する手段を有することを特徴とする請求項１記載の情報処理システム。

【請求項３】

前記転送手段は、

前記複数のディスク装置のうちの所定のディスク装置に前記情報を書き込む命令を発行する発行手段を含み、

前記記憶装置は、

前記発行手段によって発行された命令によって書き込まれた前記情報を当該記憶装置が読み出した場合に、前記読み出し手段及び前記複写手段を実行することを特徴とする請求項１又は２記載の情報処理システム。

【請求項４】

前記記憶装置は、

前記複写手段によって複写されたデータに対してアクセスがある場合には、当該アクセスがあったことを記録するアクセス記録手段と、

前記アクセス記録手段に記録された内容に基づいて、移動元のデータと移動先のデータとのデータの内容を一致させる手段とを有することを特徴とする請求項１、２又は３記載の情報処理システム。

【請求項５】

複数のディスク装置を有する記憶装置と接続される接続部と、

前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録されるメモリと、

中央演算部とを有し、

前記中央演算部は、

前記メモリから、データの移動先となる前記論理的な記憶領域に対応する第一のディスク装置の物理的な記憶領域を示す情報を取得する取得手段と、

前記取得手段によって取得された前記第一のディスク装置の物理的な記憶領域を示す情報及び前記データが格納されている第二のディスク装置の物理的な記憶領域を示す情報を、前記接続部を介して前記記憶装置に転送する転送手段と、

前記データの移動が終了した後、前記メモリに登録された前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係を更新する手段とを有することを特徴とする情報処理装置。

【請求項６】

前記取得手段は、

前記メモリから論理的な記憶領域が割り当てられていない前記複数のディスク装置の物理的な記憶領域を検索する手段と、

前記検索する手段によって検索された前記物理的な記憶領域を前記第一のディスク装置の物理的な記憶領域として取得する手段とを含むことを特徴とする請求項５記載の情報処理装置。

【請求項７】

前記転送手段は、

前記記憶装置が有する複数のディスク装置のうち、所定のディスク装置に、前記情報をデータとして書き込む命令を発行するものであることを特徴とする請求項６記載の情報処理装置。

【請求項８】

ホストコンピュータに接続される接続部と、

複数の記憶領域と、

前記複数の記憶領域のうち、前記ホストコンピュータに使用されていない記憶領域の情報が登録されるメモリと、

前記複数の記憶領域へのデータの入出力を制御する制御部を有し、

前記制御部は、

前記接続部から入力される情報に基づいて、前記メモリに登録された記憶領域のうちのいずれかを選択する手段と、

前記選択された記憶領域に、他の前記記憶領域からデータを複写する複写手段とを有することを特徴とする記憶装置。

【請求項９】

前記接続部から入力される情報とは、所定の条件を満たすことを要求する情報であり、

前記メモリには、複数の記憶領域の性質が登録され、

前記制御部は、

前記要求を満たす記憶領域を、前記メモリに登録された前記記憶装置の性質に基

づいて検索する検索手段と、前記検索された記憶領域についての情報を前記接続部から出力する出力手段を有することを特徴とする請求項８記載の記憶装置。

【請求項１０】

ホストコンピュータ及び前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置を有する情報処理システムにおいて、前記複数のディスク装置内でデータを再配置する方法であって、

前記ホストコンピュータにおいて、

再配置の対象となるデータが格納されている第一のディスク装置を特定し、

再配置先となる第二のディスク装置の情報を取得し、

特定した前記第一のディスク装置の情報及び取得した前記第二のディスク装置の情報を前記記憶装置に送信し、

前記記憶装置において、

送信された前記第一のディスク装置の情報に基づいて前記第一のディスク装置からデータを読み出し、

送信された前記第二のディスク装置の情報に基づいて前記第二のディスク装置に前記第一のディスク装置から読み出したデータを格納し、

前記読み出されたデータの格納が終了したら、前記ホストコンピュータに格納の完了を通知し、

前記通知の後、前記ホストコンピュータにおいて、前記複数のディスク装置と論理的な記憶領域との対応関係を記録したテーブルを変更することを特徴とするデータの再配置の方法。

【請求項１１】

ホストコンピュータ及び前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置を有する情報処理システムにおいて、前記複数のディスク装置内でデータを再配置するコンピュータプログラムであって、

前記ホストコンピュータにおいて、

再配置の対象となるデータが格納されている第一のディスク装置を特定し、

再配置先となる第二のディスク装置の情報を取得し、

特定した前記第一のディスク装置の情報及び取得した前記第二のディスク装置の

情報を前記記憶装置に送信するプログラムと、

前記記憶装置において、

送信された前記第一のディスク装置の情報に基づいて前記第一のディスク装置からデータを読み出し、

送信された前記第二のディスク装置の情報に基づいて前記第二のディスク装置に前記第一のディスク装置から読み出したデータを格納し、

前記読み出されたデータの格納が終了したら、前記ホストコンピュータに格納の完了を通知するプログラムとから構成され、

前記ホストコンピュータ側のプログラムは、前記通知の後、前記複数のディスク装置と論理的な記憶領域との対応関係を記録したテーブルを変更することを特徴とするデータの再配置を行うコンピュータプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ホストコンピュータ（ホスト）及び該ホストと接続される記憶装置を有する計算機システムに係り、特に、計算機システムが有する記憶装置に格納されたデータの移動を支援する機能に関する。

【０００２】

【従来の技術】

一般に、計算機システムを構築する際には、ネットワークやディスク装置などのリソースがボトルネックとならないようにシステム設計が行われる。特に、プロセッサなどに比べて低速な外部記憶装置は、性能上のボトルネックとなりやすく、システム設計にあたって様々な対策が施される。その一つとして、記憶装置上のデータ格納形式の最適化が挙げられる。

例えば、高頻度でアクセスされるデータをより高速なディスク装置に格納したり、データを複数のディスク装置に分散配置することでデータアクセスの性能を改善できる。また、いわゆるＲＡＩＤ(Redundant Array of Independent Disk)装置を使用する場合には、ＲＡＩＤ装置のデータアクセスのシーケンシャル性を考慮して、ＲＡＩＤレベル(冗長構成)を決定すれば、アクセス特性に適合したデー

タの格納が可能となる。

【０００３】

システム設計という視点から見た場合には、システムにおいて使用されるデータの格納形式を考慮して、各データに割り当てられるディスク装置の容量を決定することが必要となる。具体的には、データベース(ＤＢ)におけるＤＢ表の領域サイズや、ファイルシステム(ＦＳ)におけるＦＳサイズの決定がこれに相当する。通常、計算機システムの運用に伴って取り扱われるデータは増加していく。このため、システム設計時には、関連業務における過去の実績などを元にデータ量の増加率を予測し、次に保守可能な時期までに予測されるデータ量の増加に耐えられるだけの空き領域を確保するようにディスク装置の容量を決定して、それぞれのデータ領域が定義される。

【０００４】

上述のように、システム設計に際しては、データアクセス性能の向上およびデータ量の増加を考慮して、記憶装置とデータの格納形式を組み合わせることが重要となる。この組み合わせの決定を支援する手段として、Logical Volume Manager（ＬＶＭ）がある。

【０００５】

ＬＶＭとは、実際のディスク装置の任意の部分領域をまとめ、論理的に一つのボリューム（これを論理ボリュームと言い、以下「ＬＶ」と称する。）としてホストに提供するものである。ＬＶＭは、ＬＶを管理し、ＬＶの作成、削除、サイズ変更(拡大/縮小)なども行う。ＬＶＭは、ＬＶを多重化するミラーリング機能や、ＬＶを小領域(ストライプ)に分割して、複数の物理ボリューム(ＰＶ)に分散して配置するストライピング機能も有する。

【０００６】

ＬＶＭを用いた場合、ユーザは、ＤＢ表やＦＳといったデータを格納する領域をＰＶではなくＬＶ上に配置することになる。このようにすることで、データの格納形式の選択又は管理が容易となる。例えば、ＬＶ上にＦＳを配置することで、通常１台のディスク装置、またはその１つのパーティションに１つしか割り当てられないＦＳを、複数のディスク装置に分散して配置することが可能となる。

また、ファイルの容量の増加に合わせてＬＶを拡張することで、最小限の手間でＦＳを拡張(再構成)することが可能となる。

【０００７】

【発明が解決しようとする課題】

計算機システムで業務運用を続けていくと、データの格納形式を見直す必要が生じることがある。見直しの契機としては、データのアクセス傾向やアクセス特性の変化、当初見積もりとは異なるデータ量の変動などといった、システム設計時に想定されていた業務モデルの見直しによるもの、ディスク装置の追加や高速なリソースへの置き換えなどといった、前提とする物理リソース構成の変更によるもの、拡張を繰り返すことによる断片化(fragmentation)したＬＶやＤＢ表の連続化などといったデータ管理方式上必然的に生じるものなどが考えられる。このような場合、データの格納形式を見直し、データを再配置することで、システムの性能を改善することができる。

【０００８】

しかし、記憶装置に格納されたデータの再配置を行うには、従来の技術においては、ホストの介在によるデータ転送が必要となる。

【０００９】

今、複数のＰＶに分散されたＬＶを一つのＰＶにまとめる場合のデータ再配置処理の手順を以下に示す。

(1)処理対象となるＬＶの容量分の領域をＰＶ上に確保する。

(2)ＬＶからホストがデータをリードし、それを新しいＬＶの領域にライトする。

(3)(2)を繰り返して全データをコピーし終えたら、ＬＶとＰＶマッピング情報を変更する。

【００１０】

このように、ＬＶを再構成するには、ＬＶ全体という大量のデータ転送が生じるため、データ転送元/先のＰＶに大量の入出力（I/O）が発生する。同時に、ホストおよびチャネルにも多大な負荷がかかり、他のＬＶ内のデータを対象とした実行中の処理にも性能上悪影響を及ぼしてしまう。

【００１１】

また、このようなデータ再配置処理の間は、再配置の対象となるデータへのアクセスについて、少なくともデータの更新を抑止する必要がある。例えば、ＬＶの再構成の場合、通常、ＬＶをオフライン(Unixではアンマウント)とし、再構成完了後にオンライン(Unixではマウント)にすることで、データ再配置中のデータのアクセスを抑止する。データアクセスの対象となるＬＶがオフラインとなるため、そのＬＶを利用する業務は、データ再配置処理の間、停止することになる。

【００１２】

このため、ＬＶ再構成等のデータ再配置を行う場合には、そのデータにアクセスする業務を所定の期間中断することができる時間帯に行わなければならない。従って、計算機システムの保守作業に時間的制約が生じてしまう。

【００１３】

本発明の目的は、記憶装置に格納されたデータを再配置する際に、移動元データ領域から移動先データ領域へのデータ転送によるホストおよびチャネルの負荷を軽減することにある。

【００１４】

また、本発明の別の目的は、データ再配置により、データへのアクセスが不可となる期間を可能な限り削減し、そのデータを利用する業務が中断される時間を削減することにある。

【００１５】

【課題を解決するための手段】

上記目的を達成するために、本発明では、ホストコンピュータと、前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置とを有する情報処理システムで、以下の構成とする。ホストコンピュータは、複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録されたテーブルを有する。ホストコンピュータは、第一のディスク装置に記録されたデータを第二のディスク装置に移動する場合に、第二のディスク装置の物理的な記憶領域を示す情報をテーブルから取得する。ホストコンピュータは、取得した情報を、記憶装置に転送する。データの移動が終了した後、ホストコンピュータは、テ

ーブルに登録された情報を更新する。また、記憶装置は、ホストコンピュータから転送された前記情報を用いて、移動元のデータを移動先のディスク装置に複写する。

【００１６】

以上の構成により、データを再配置する場合のデータの転送を記憶装置側で行うことで、データ転送によるホストおよびチャネル負荷を軽減することができる。

【００１７】

また、記憶装置でのデータ転送処理を、単純に領域間コピーで実現するのではなく、転送元と転送先のデータ領域で一時的に内容の同期をとるペアを形成し、データ転送中およびデータ転送完了後に転送元データ領域に対して行われたデータ更新を全て転送先へ反映する構成とすることも出来る。

【００１８】

また、ホストコンピュータではなく、記憶装置でディスク装置の仮想的な記憶領域の管理を行う構成も考えられる。

【００１９】

【発明の実施の形態】

図１は、本発明が適用された計算機システムの第１実施形態における構成を示す構成図である。

【００２０】

本実施形態における計算機システムは、ホスト１００及び記憶装置１１０を有する。ホスト１００と記憶装置１１０は、ＳＣＳＩバス等の通信線で接続されている。両者間の交信は通信線を介して行われる。

【００２１】

ホスト１００は、ＣＰＵ１０１、主記憶１０２及びチャネル１０３を有し、それぞれが内部バスで接続される。

【００２２】

ＣＰＵ１０１は、データベース等のアプリケーションプログラム（ＡＰ）を実行する。ＡＰが動作する際のメモリ割り当てや、記憶装置１１０への入出力制御

は、ＣＰＵ１０１で動作するオペレーティングシステム（ＯＳ）およびＯＳに関連するソフトウェアで行われる。ＬＶＭ１４２も、ＯＳに関連するソフトウェアである。ＬＶＭ１４２は、記憶装置１１０により提供されるＰＶの記憶領域をまとめてＬＶとして仮想的にＡＰに提供する。

本実施形態では、ＣＰＵ１０１で、ＬＶＭ１４２が実行される。ＬＶＭ１４２は、後述するＬＶ－ＰＶ対応情報１４１などの制御情報を用いてＬＶの制御を行う。

【００２３】

主記憶１０２には、アプリケーションプログラム、ＯＳ、ＬＶＭ等のＯＳ関連ソフトウェアの実行オブジェクトコードや各ソフトウェアが使用するデータや制御情報等が格納される。

【００２４】

図３及び図４は、ＬＶ－ＰＶ対応情報１４１のテーブル図である。

ＬＶ－ＰＶ対応情報１４１は、ＬＶが対応するＰＶ(又はＰＶが対応するＬＶ)を示す情報である。ＬＶ－ＰＶ対応情報１４１のテーブルには、ＬＶ管理情報３００とＰＶ管理情報４００が、１つのＬＶ（ＰＶ）ごとに設けられる。

【００２５】

ＬＶ管理情報３００は、ＰＶリスト３０１、ＬＥ数３０２、ＬＥサイズ３０３及びＬＥ－ＰＥ対応情報３１０のエントリを有する。

ＰＶリスト３０１には、ＬＶに対して割り当てられているＰＶを示す情報が登録される。ＬＶとＰＶは、それぞれＬＥ(Logical Extent)およびＰＥ(Physical Extent)という同じサイズの小領域に分割されている。ＬＥに対してＰＥを割り当てることで、ＬＶの物理配置の自由度が向上する。ＬＥ数３０２には、ＬＶに含まれるＬＥ数が登録される。ＬＥサイズ３０３には、ＬＥのサイズを示す情報が登録される。ＬＥ－ＰＥ対応情報３１０のエントリは、ＬＥ番号３１１に割り当てられるＬＥに対応するＰＶ名３１２およびＰＥ番号３１３のエントリを有する。ＬＥ－ＰＥ対応情報３１０には、ＬＥが割り当てられているＰＥを示す情報が登録される。

【００２６】

ＰＶ管理情報４００は、ＬＶ管理情報とは逆に、各ＰＶに割り当てられるＬＶの情報を示すものである。

【００２７】

ＬＶ管理情報４００は、ＬＶリスト４０１、ＰＥ数４０２、ＰＥサイズ４０３及びＰＥ－ＬＥ対応情報４１０のエントリを有する。
ＬＶリスト４０１には、ＰＶに対して割り当てられているＬＶを示す情報が登録される。ＰＥ数４０２には、ＰＶに含まれるＰＥ数が登録される。ＰＥサイズ４０３には、ＰＥのサイズを示す情報が登録される。ＰＥ－ＬＥ対応情報４１０は、ＰＥ番号４１１に割り当てられるＰＥに対応するＬＶ名４１２およびＬＥ番号４１３のエントリを有する。ＰＥ－ＬＥ対応情報４１０には、ＰＥが割り当てられているＬＥを示す情報が登録される。

【００２８】

主記憶１０２には、上記した情報以外にも、ＰＶへアクセスするために必要な情報が保持される。例えば、ＰＶへアクセスするための経路情報として、接続チャネル１０３の番号および記憶装置１１０のポート１１４の番号や、記憶装置１１０内でのデバイス番号(以下、ＰＶ番号と呼ぶ)が格納される。

【００２９】

チャネル１０３は、通信線を介して記憶装置１１０への入出力の制御を行うコントローラである。チャネル１０３は、通信線上の要求コマンドの送信、完了報告の通知、対象データの送受信、通信フェーズ管理等、通信プロトコルの制御を行う。通信線がＳＣＳＩバスの場合、ＳＣＳＩアダプタカードが、チャネル１０３に相当する。

【００３０】

記憶装置１１０は、ホストとの接続の制御を行うポート１１４、ディスク装置１５０、記憶制御プロセッサ１１１、制御メモリ１１２及びディスクキャッシュ１１３を有する。記憶装置１１０の可用性を向上させるため、記憶装置１１０の各部位は多重化されている。したがって、ある部位に障害が発生した場合にも、残った正常な構成部位による縮退状態での運用が可能である。

【００３１】

記憶装置１１０が、ディスク装置１５０が複数台接続されたＲＡＩＤシステムの場合、記憶制御プロセッサ１１１による論理物理対応管理によるエミュレーションを使用することによって、複数台のディスク装置１５０を、論理的に１台又は複数のディスク装置としてホスト１００に認識させることができる。ただし、本実施形態では、説明を簡略化するため、ホスト１００がアクセスするＰＶ、すなわち記憶装置１１０での論理ディスク装置は、ディスク装置１５０と一対一に対応しているものと仮定する。

【００３２】

記憶制御プロセッサ１１１は、ホスト１００からＰＶへのアクセスを受け付け、ディスク装置１５０とディスクキャッシュ１１３間のデータ転送、ディスクキャッシュ１１３とホスト１００間のデータ転送、記憶装置１１０内でのディスク装置１５０の論理物理対応の管理、ディスクキャッシュ１１３の領域管理などを行う。

【００３３】

ディスクキャッシュ１１３には、ホスト１００からのライトデータや、ディスク装置１５０からのリードデータが、転送先に送信されるまでの間、一時的に保持される。ディスクキャッシュ１１３に格納されたデータは、ＬＲＵ(Least Recently Used)などの方式で管理される。ディスクキャッシュ１１３を用いて、ライトデータをホストのＩ／Ｏ要求とは非同期にディスク装置１５０へ書き込む処理も行われる。これらのキャッシュの制御方法は従来公知の技術であるので、説明は省略する。

【００３４】

制御メモリ１１２には、ディスク装置１５０への入出力を制御するために記憶制御プロセッサ１１１が用いる各種制御情報のテーブルが格納される。制御情報のテーブルには、ディスクキャッシュ１１３の領域割り当て管理に用いるキャッシュ管理情報１４４、論理ディスク装置とディスク装置１５０との対応等の管理に用いられる記憶装置管理情報１４３、ホスト１００から指示された領域間のデータ転送処理の対象範囲や処理進捗状況等の管理に用いられるデータ転送領域情報１４５がある。

【００３５】

図５は、データ転送領域情報１４５のテーブル図である。

転送元範囲情報５０１及び転送先範囲情報５０２には、ホスト１００から指示されたデータ転送の対象となるデータ領域の範囲を示す情報が登録される。本実施形態では、転送元/先のデータ領域が不連続であるケースを想定し、転送元範囲情報５０１及び転送先範囲情報５０２には、連続する部分領域毎に、その部分領域を含むＰＶ番号、ＰＶ内相対アドレスで示される先頭位置およびサイズを連ねたリストが登録される。尚、転送元/先のデータ領域の総容量は等しくなければならない。

【００３６】

進捗ポインタ５０３には、データ転送処理で転送されたデータの量を示す情報が登録される。進捗ポインタ５０３に登録された情報を用いて、データ転送処理の進捗が管理される。同期状態５０４と差分ビットマップ５０５は、本実施形態では使用されないので、説明を省略する。

【００３７】

本実施形態でのＣＰＵ１０１と記憶制御プロセッサ１１１の動作を説明する。

【００３８】

ユーザ又は保守員は、各ＬＶのＰＶへの割り当て状態等の情報から、特定のＬＶについて再構成が必要であることを判断すると、ＬＶの再構成の指示を行う。本実施形態では、図２に示すように、lv0が、２台のＰＶ、pv0とpv1上にlv0_0とlv0_1として格納されているとき、lv0をpv2内領域に新規に確保されたlv0_0’とlv0_1’に移動する指示を出した場合について説明する。

【００３９】

ＬＶの再構成は、ＣＰＵ１０１で動作するデータ再配置処理１３１及び記憶装置１１０の記憶制御プロセッサ１１１で動作するコマンド処理１３２が協働することで実現される。

【００４０】

図６は、ＣＰＵ１０１が行うデータ再配置処理１３１のフロー図である。

データ再配置処理１３１は、ユーザ等がＬＶの再配置を指示した場合に実行され

る。データ再配置処理１３１では、その処理が実行される前に、再配置対象であるＬＶ名と再配置先のＰＶ名が取得される。

【００４１】

ＣＰＵ１０１は、処理対象となるＬＶに対するアクセスを抑止するために、対象となるＬＶをオフラインにする。ＬＶのオフラインは、例えば、ＯＳがＵＮＩＸである場合、デバイス(ＬＶ)をアンマウントすることで実現できる（ステップ６０１）。

【００４２】

ＣＰＵ１０１は、ＬＶ－ＰＶ対応情報１４１のＬＶ管理情報３００を参照して、対象となるＬＶのデータが格納されているＰＶおよびＰＥを求める。ＣＰＵ１０１は、ＬＥ数３０２とＬＥサイズ３０３から、対象となるＬＶサイズを算出する。対象となるＬＶの全体もしくは一部が既に転送先ＰＶに格納されている場合には、ＣＰＵ１０１は、転送先ＰＶに格納されている部分については、転送を行わない。ただし、以下の説明では、転送対象からはずされる部分は存在しないと仮定する（ステップ６０２）。

【００４３】

ＣＰＵ１０１は、対象となるＬＶの転送対象領域サイズ分のＰＥを、転送先ＰＶ上に確保する。具体的に、ＣＰＵ１０１は、ＰＶ管理情報４００のＰＥ－ＬＥ対応情報４１０を参照して、ＬＥに未割り当てのＰＥを求め、転送先として必要となる容量分のＰＥを確保する。ＰＥの確保は、転送先ＰＶおよびＰＥが記憶されるだけでもよい。しかし、他の処理によりＰＥを別の目的で確保されてしまう可能性がある場合は、転送に使用するＰＥをあらかじめ確保しておく。具体的には、ＰＶのＰＥ割り当ての変更を一定期間禁止することを示す構成変更処理排他フラグをＰＶ毎に設ける、構成変更処理排他フラグをＰＶ毎ではなくＰＥ毎に設ける、又はデータ転送が完了する前に、ＰＥ－ＬＥ対応情報４１０の対象ＰＥ項目を、転送元ＬＶに割り当て済みであるように変更する等の方法が考えられる（ステップ６０３）。

【００４４】

転送先のＰＶを確保したら、ＣＰＵ１０１は、転送元のＰＶ領域をいくつかの

部分領域に分割し、各部分領域毎のデータ転送処理要求を記憶装置１１０に発行する。記憶装置１１０へのデータ転送処理の要求は、既存のプロトコルで準備されている標準の入出力コマンドではなく、データ転送処理の要求用に新たに追加した専用のコマンドを用いて行われる。ＰＶ領域の分割は、適当なサイズで転送元領域を先頭から順に区切って行われる。分割に適当なサイズは、記憶装置１１０へ要求されたデータの転送処理に要する時間と、要求元ホスト１００でのその要求に対する応答時間として許容可能な時間とから決定される。転送処理要求のコマンドがＰＶ毎に発行されるので、転送元ＬＶが複数のＰＶ上に分散して配置されている場合には、部分領域が２つのＰＶにまたがらないように分割する必要がある。転送処理要求のコマンドには、転送元ＰＶの先頭アドレスとサイズ、および転送先のＰＶ番号、ＰＶ内先頭アドレス、データサイズが含まれる。チャネル１０３を介してデータ転送要求コマンドを記憶装置１１０へ送出したら、ＣＰＵ１０１は、記憶装置１１０からの完了報告を待つ（ステップ６０４）。

【００４５】

ステップ６０４で送信したデータ転送処理要求コマンドの完了報告を受信したＣＰＵ１０１は、転送元ＬＶについて全対象領域のデータ転送が完了したかをチェックする。データ転送が完了していない部分領域があれば、ＣＰＵ１０１は、ステップ６０４の処理に戻る（ステップ６０５）。

【００４６】

対象領域の全てのデータが転送されたら、ＣＰＵ１０１は、転送の対象となったＬＶが転送先ＰＶへ対応づけられるよう、ＬＶ－ＰＶ対応情報１４１を変更する。具体的には、ＣＰＵ１０１は、ＬＶ管理情報３００のＰＶリスト３０１に登録された情報を転送先ＰＶを示す情報に変更し、ＬＥ－ＰＥ対応情報３１０のＰＶ名３１２とＰＥ番号３１３に登録された情報を、転送先ＰＶのＰＥを示す情報となるよう変更する。ＣＰＵ１０１は、転送先ＰＶのＰＶ管理情報４００のＬＶリスト４０１に、転送対象となったＬＶを追加し、ＰＥ－ＬＥ対応情報４１０のＬＶ名４１２とＬＥ４１３番号を、転送先ＬＶのＬＥに対応するよう変更する。ＣＰＵ１０１は、転送元ＰＶのＰＶ管理情報４００のＬＶリスト４０１から転送対象となったＬＶを削除し、ＰＥ－ＬＥ対応情報４１０の転送元ＰＥ項目をＬＶ

未割り当てに変更する（ステップ６０６）。

【００４７】

その後、ＣＰＵ１０１は、転送対象のＬＶのオフラインを解除し、処理を終了する（ステップ６０７）。

【００４８】

図７は、記憶装置１１０で実行される、コマンド処理１３２のフロー図である。コマンド処理１３２は、ホスト１００のコマンドを記憶装置１１０が受領した時に実行される。

【００４９】

記憶装置１１０は、ホスト１００からディスク装置１５０に発行された処理要求コマンドの種別をチェックする（ステップ７０１）。

【００５０】

コマンドがデータ転送処理要求であった場合、記憶装置１１０は、コピー処理１３３を起動し、その完了を待つ（ステップ７０２）。

【００５１】

コマンドがリード要求であった場合、記憶装置１１０は、そのリード要求に対して、対象となるデータがディスクキャッシュ１１３上に有るか否かをチェックする。必要ならば、記憶装置１１０は、キャッシュ領域を確保して、ディスク装置１５０から対象となるデータを、確保したキャッシュ領域へ読み出し、ホスト１００へ転送する（ステップ７０３～ステップ７０７）。

【００５２】

コマンドがライト要求であった場合、記憶装置１１０は、ディスクキャッシュ１１３のキャッシュ領域を確保し、ホスト１００から受け取ったデータを一旦そのキャッシュ領域に書き込む。その後、ディスク装置１５０へライトする（ステップ７０９～ステップ７１２）。

【００５３】

記憶装置１１０は、要求処理の完了をホスト１００へ報告し、処理を終了する（ステップ７０８）。

【００５４】

図８は、記憶装置１１０が行うコピー処理１３３のフロー図である。コピー処理１３３は、記憶装置１１０がデータ転送要求のコマンドを受けとったときに実行される。

【００５５】

データ転送要求を受け取った記憶装置１１０は、そのデータ転送要求について転送元/先データ領域の指定情報の妥当性をチェックする。
具体的には、転送元/先データ領域のサイズが同じであるか、既に別のデータ転送要求の転送元/先データ領域として設定されていないか等をチェックする。指定情報が妥当でない場合、記憶装置１１０は、エラーをホスト１００へ報告する（ステップ８０１）。

【００５６】

エラーが発見されない場合、記憶装置１１０は、データ転送処理要求に対応するデータ転送領域情報１４５を格納する領域を制御メモリ１１２上に割り当て、初期化する。具体的には、転送元範囲情報５０１と転送先範囲情報５０２を、受け取ったデータ転送処理要求に含まれる情報を元に設定し、進捗ポインタ５０３を初期値である０に設定する（ステップ８０２）。

【００５７】

設定が済んだら、記憶装置１１０は、転送対象となる転送元ディスク装置１５０のデータ領域の先頭から、順次データをディスクキャッシュ１１３にリードし、リードしたデータを転送先ディスク装置１５０へライトする。このときの一回のデータ転送のデータ量は、ディスク装置１５０におけるヘッド位置づけオーバヘッドを考慮すると、ある程度大きいサイズのデータ領域であることが望ましい。しかし、1回のデータ量が大きすぎると、ディスク装置１５０と同じバスに接続される他のディスク装置１５０に格納されたデータにアクセスする他の処理に悪影響を及ぼす可能性がある。したがって、一回のデータ転送のデータ量は、コピー処理１３３に期待される処理速度と、他の処理への影響とを考慮して決定する必要がある(ステップ８０３～８０４)。

【００５８】

ステップ８０４で転送先への書き込みを終えたら、記憶装置１１０は、転送し

たデータ容量に応じて、進捗ポインタ５０３の内容を更新する。

【００５９】

記憶装置１１０は、進捗ポインタ５０３を参照して、全対象データについてコピー処理が完了したかをチェックする。コピー処理が完了していない部分があれば、ステップ８０４の処理へ戻る（ステップ８０５）。

【００６０】

コピー処理が完了してれば、記憶装置１１０は、コピー処理の完了を、コマンド処理１３２へ報告し、処理を終了する（ステップ８０６）。

【００６１】

本実施形態により、ホスト１００はコピー処理の指示だけを行えば良く、実際のデータ転送は記憶装置が行うので、ホストやネットワーク等にかかる負荷が軽減されることとなる。

【００６２】

図９は、本発明の第２実施形態が適用された計算機システムの構成図である。本実施形態では、データ転送領域情報１４５に含まれる同期状態５０４及び差分ビットマップが使用される点及びコマンドボリューム９００を有する点が、第１実施形態と異なる。以下、第２実施形態に特有な部分のみを説明する。

【００６３】

同期状態５０４には、データ転送処理の転送先/元データ領域の同期ペア状態を示す情報が登録される。同期ペア状態の取りうる値としては、「ペア未形成」、「ペア形成中」、「ペア形成済」がある。「ペア形成中」の状態とは、転送が指示されたデータ領域間で、転送元から転送先へデータの転送処理が実行中である状態を示す。「ペア形成済」の状態とは、データ領域間でのコピー処理が完了し、同期ペアが形成されていることを示す。ただし、「ペア形成中」の状態において転送元データ領域のデータの更新があった場合、「ペア形成済」状態でも、同期ペアのデータ領域間に差分がある可能性がある。「ペア未形成」の状態とは、データ領域間のデータ転送が指示されていないか、データ転送が完了した後、ホスト１００の指示により同期ペアが解除された状態であることを示す。ただし、この状態は、データ転送処理がそもそも存在しないか、完了してしまっている

状態なので、制御メモリ１１２上には、データ転送領域情報１４５は確保されない。したがって、実際に同期状態５０４に設定されるのは、「ペア形成中」と「ペア形成済」の２つである。

【００６４】

差分ビットマップ５０５は、ペア形成中およびペア形成済の状態において、コピー元データ領域内のデータ更新の有無を示す。情報量を削減するため、ディスク装置１５０の全てのデータ領域を、例えば64KBの特定サイズの小領域に分割し、その小領域単位と差分ビットマップ５０５の１ビットを一対一に対応させる。差分ビットマップ５０５には、小領域に対するデータの更新の有無が記録される。

【００６５】

ディスクキャッシュの割り当てにおいても、同様にディスク装置１５０を小領域に分割し、小領域毎にキャッシュを割り当てることで、キャッシュ管理を簡単化することが多い。この場合、差分ビットマップ５０５の１ビットをキャッシュ割り当て単位である小領域１つもしくは複数に対応づけることで、ビットマップの設定を容易にできる。

【００６６】

コマンドボリューム９００には、標準のプロトコルには含まれない特殊な処理要求（データ転送処理要求等）が、データとして書き込まれる。第１実施形態では、記憶装置１１０に対するデータ転送処理要求のコマンドとして専用コマンドを追加した。本実施形態では、通常のライト要求コマンドを用い、データ転送要求がデータとしてコマンドボリューム９００に書き込まれることで、記憶装置１００へデータ転送要求が発行される。

【００６７】

記憶制御プロセッサ１１１は、コマンドボリューム９００に対するライト要求を受け付けると、ライトデータを処理要求として解析し、対応する処理を起動する。要求される処理をライト要求の延長で実行しても応答時間が問題にならない程度短いならば、記憶制御プロセッサ１１１は、要求処理を実行し、ライト要求と合わせて完了報告を行う。要求される処理の実行時間がある程度長い場合には

、記憶制御プロセッサ１１１は、一旦ライト処理について完了報告を行う。ホスト１００は、周期的にその要求処理が完了したかをチェックする。

【００６８】

次に、本実施形態でのＣＰＵ１０１及び記憶制御プロセッサ１１１の動作を説明する。

【００６９】

ＣＰＵ１０１で動作するデータ再配置処理１３１及び記憶装置１１０の記憶制御プロセッサ１１１で動作するコマンド処理１３２が、協働してＬＶの再構成を行うのは、第１実施形態と同様である。

【００７０】

図１０は、本実施例のデータ再配置処理１３１のフロー図である。

【００７１】

ステップ１００１、１００２は、図６のステップ６０２、６０３と同様の処理のため、説明は省略する。

【００７２】

ＣＰＵ１０１は、ＬＶの転送対象領域に対応する全ＰＶ領域のデータ転送処理の要求を、記憶装置１１０へ一括して発行する。コマンドボリューム９００へのライトデータに含まれる転送要求には、ＬＶ対象領域に対応する全ＰＶ領域の範囲情報(各部分領域の位置情報(各ＰＶ番号、先頭アドレス、サイズ)のリスト)及び転送先ＰＶ領域の範囲情報等のパラメタが含まれる。記憶装置１１０へライト要求コマンドを発行したら、ＣＰＵ１０１は、記憶装置１１０からの完了報告を待つ（ステップ１００３）。

【００７３】

完了報告を受け取ったら、ＣＰＵ１０１は、規定の時間が経過するのを待つ（ステップ１００４）。

【００７４】

ＣＰＵ１０１は、データ転送領域の同期ペア状態を参照する要求を記憶装置１１０へ発行し、その完了を待つ。同期ペア状態を参照するには、具体的には、ＣＰＵ１０１は、同期ペア状態の準備要求を含むデータをコマンドボリューム９０

０へ書き込むためのライト要求を発行する。ＣＰＵ１０１は、記憶装置１１０から完了報告を受けた後、コマンドボリューム９００へのリード要求を発行する（ステップ１００５）。

【００７５】

ＣＰＵ１０１は、得られた同期ペア状態が「ペア形成済み」であるか否かを判定する。同期ペア状態が「ペア形成済み」である場合、ＣＰＵ１０１は、ステップ１００７の処理を行う。同期ペア状態が「ペア形成済み」ではない場合、ＣＰＵ１０１は、ステップ１００４の処理に戻り、同期ペア状態が変更されるのを待つ（ステップ１００６）。

【００７６】

その後、ステップ６０１と同様に、ＣＰＵ１０１は、ＬＶをオフラインにする（ステップ１００７）。

【００７７】

ＣＰＵ１０１は、データ転送の転送元領域と転送先領域で形成済みの同期ペアの解除要求を、コマンドボリューム９００を用いて発行する（ステップ１００８）。コマンドボリューム９００へ同期ペア解除要求を転送するライト要求コマンドの完了報告が記憶装置１１０より報告されたら、ＣＰＵ１０１は、ステップ１００４と同じ方法で、データ領域の同期ペア状態を参照する（ステップ１００９）。

【００７８】

取得された同期ペア状態が「ペア未形成」でなければ（ステップ１０１０）、ＣＰＵ１０１は、規定時間が経過するのを待ち（ステップ１０１１）、ステップ１００９の処理に戻って同期状態５０４の内容を再取得する。同期状態５０４が「ペア未形成」となっていれば、ＣＰＵ１０１は、ステップ１０１２以降の処理を行う。ＣＰＵ１０１は、ＬＶのＬＶ－ＰＶ対応情報１４１を更新し、ＬＶをオンラインにしてＬＶ再配置処理を完了する。

【００７９】

図１１は、コマンド処理１３２のフロー図である。

【００８０】

記憶装置１１０は、ホスト１００から受け取ったコマンドの対象が、コマンドボリューム９００であるか否かを判定する（ステップ１１０１）。

【００８１】

コマンドの対象がコマンドボリューム９００であれば、記憶装置１１０は、コピー処理を起動し、その完了を待つ（ステップ１１０２）。

【００８２】

コマンドの対象がコマンドボリューム９００でなければ、記憶装置１１０は、コマンドの種別を判定する。コマンドがリードかライトかでそれぞれステップ１１０４、１１１０へ遷移する（ステップ１１０３）。

コマンドの種別がリードである場合には、記憶装置１１０は、図７のステップ７０３から７０７と同様のリード処理を行う（ステップ１１０４～１１０８）。

【００８３】

コマンド種別がライトである場合には、記憶装置１１０は、図７のステップ７０９から７１２と同様のライト処理を行う（ステップ１１１０～１１１３）。

【００８４】

ステップ１１１４、１１１５は、本実施形態に特有な部分で、データ転送処理中にホスト１００の別の処理が転送中のＬＶのデータに対してアクセスする場合の処理部分である。

【００８５】

記憶装置１１０は、ライトの対象となったデータに、データ転送領域として登録されているデータ領域が含まれていないかを判定する（ステップ１１１４）。登録されているデータ領域が含まれている場合、記憶装置１１０は、データ転送領域のうち、更新された部分を特定し、更新部分に対応した差分ビットマップを設定する（ステップ１１１５）。

【００８６】

記憶装置１１０は、要求処理の完了をホスト１００へ報告する（ステップ１１０９）。

【００８７】

図１２は、コピー処理１３３のフロー図である。

【００８８】

記憶装置１１０は、コマンドボリューム９００へ送信されたコマンドの種別を切り分ける（ステップ１２０１）。

【００８９】

コマンドの種別がライトの場合、記憶装置１１０は、コマンドボリューム９００へのライトデータの内容を解析し、処理要求内容や、転送元/先データ領域範囲指定の妥当性を判定する（ステップ１２０２）。妥当性に問題のある場合は、記憶装置１１０は、エラーを上位処理へ報告して処理を中断する（ステップ１２０３）。

問題ない場合は正常終了を報告し、記憶装置１１０は、ライトデータとして送られてきた処理要求の種別を判定する（ステップ１２０４）。

【００９０】

処理要求がデータ転送の場合、記憶装置１１０は、図８のステップ８０２から８０５と同様のデータ転送処理を行う（ステップ１２０５～１２０８）。ただし、ステップ１２０５におけるデータ転送領域情報１４５の初期化処理では、同期状態５０４を「ペア形成中」にし、差分ビットマップ５０５を全て０クリアする処理を行う。データ転送処理が完了したら、記憶装置１１０は、同期状態５０４を「ペア形成済み」に変更し、処理を終了する（ステップ１２０９）。

【００９１】

処理要求がペア解除の場合、記憶装置１１０は、解除対象のデータ転送領域ペアのデータ転送領域情報１４５を参照し、差分ビットマップ５０５に0nが設定されている部分がないかをチェックする（ステップ１２１０）。差分ビットマップ５０５に0であるビットが存在する、すなわちデータ転送領域に転送元/先で同期していない部分がある場合、未反映データを転送先データ領域へ転送する（ステップ１２１２）。記憶装置１１０は、ステップ１２１０へ戻り、差分ビットマップ５０５のチェックをやり直す。

【００９２】

転送元/先のデータ転送領域で同期が取れたら、記憶装置１１０は、データ領域のデータ転送領域情報１４５をクリアし、データ転送領域情報１４５を格納す

る制御メモリ１１２の領域の割り当てを解除し、処理を終了する（ステップ１２１３）。

【００９３】

処理要求がペア状態準備である場合には、記憶装置１１０は、要求されたデータ転送領域の同期ペア状態を準備する（ステップ１２１４）。データ転送領域がまだ存在し、制御メモリ１１２上にメモリ領域が割り当てられていれば、同期ペア状態として、同期状態５０４の値を用いる。データ転送領域が存在せず、メモリ領域が割り当てられていなければ、同期ペア状態として、「ペア未形成」を用いる。ステップ１２１４で準備された同期ペア状態は、コマンドボリューム９００に対して発行されたリード要求のリードデータとして転送される(ステップ１２１５、１２１６)。

【００９４】

本実施例に拠れば、第一実施例よりもＬＶをオフラインにする時間を短くした上でＬＶの再配置をホスト等に負担を掛けないで行うことが出来る。

【００９５】

第３実施形態について説明する。

【００９６】

第３実施形態のシステム構成は、基本的に第１及び第２の実施形態と同一である。ただし、本実施形態では、記憶装置１１０が、ホスト１００で使用されていないディスク装置１５０の記憶領域を管理する。そして、記憶装置１１０が、転送先のＰＶとして使用されるデータ領域の条件、例えば、領域長、領域が格納される論理ユニット番号、接続に用いる接続ポート１１４の番号およびディスクタイプなどの指示をユーザ又は保守員から受ける。記憶装置１１０は、ユーザの指示に合致するデータ領域をホスト１００が使用していないデータ領域から選びだし、ユーザ又は保守員に提示する。この点で、第１、第２実施形態と異なる。以下、記憶装置１１０における情報の提示の方法について説明する。

【００９７】

記憶装置１１０は、ホスト１００が使用していないディスク装置１５０の番号のリストを、未使用領域管理情報として保持・管理する。

【００９８】

記憶装置１１０は、ユーザまたは保守員からの領域確保の指示を受け取ると、以下の処理を行う。

【００９９】

領域確保の指示にしたがい、記憶装置１１０は、保持している未使用領域管理情報を検索して条件を満たす未使用ディスク装置１５０を選定する（ステップ１－１）。

【０１００】

記憶装置１１０は、選定したディスク装置１５０番号をユーザまたは保守員へ報告する（ステップ１－２）。

【０１０１】

ユーザ又は保守員は、記憶装置１１０から未使用のディスク装置１５０の番号を得ると、以下の手順にしたがって、データ転送の指示を行う。

ユーザ又は保守員は、未使用のディスク装置１５０のＯＳ管理情報を設定する。たとえば、ＵＮＩＸ ＯＳでは、未使用のディスク装置１５０に対して、デバイスファイル名を定義する（ステップ２－１）。

【０１０２】

ユーザ又は保守員は、ＯＳ管理情報が設定されたディスク装置１５０について、ＬＶＭが使用することができるように、ＰＶとして定義する（ステップ２－２）。

【０１０３】

ユーザ又は保守員は、新しく定義したＰＶを転送先として指定して、記憶装置１１０に対して本発明によるデータ転送を指示する（ステップ２－３）。

【０１０４】

データ転送が完了したら、ユーザ又は保守員は、ＬＶ－ＰＶ対応情報１４１の更新を指示する（ステップ２－４）。

【０１０５】

本実施形態が適用されたシステムが、例えばＲＡＩＤ装置などのように、複数のディスク装置１５０の記憶領域の一部又は全体の集合から構成する論理ディス

ク装置をホスト１００へ提示するような記憶装置１１０である場合を考える。この場合、未使用領域管理情報は、各未使用領域が属するディスク装置１５０番号および先頭領域と領域長のリストで構成される。

【０１０６】

記憶装置１１０は、次の手順で未使用領域を確保する。

【０１０７】

記憶装置１１０は、領域確保の指示の条件に合ったディスク装置１５０の未使用領域を検索する（ステップ３－１）。

【０１０８】

指示分の容量を確保できない場合は、記憶装置１１０は、確保できない旨をユーザに通知し、処理を終了する（ステップ３－２）。

【０１０９】

指示条件を満たすディスク装置１５０を一つ見つけたら、記憶装置１１０は、ディスク装置１５０に属する未使用領域の容量を確認し、指示分の領域があれば、指示分の領域を確保する。

具体的には、記憶装置１１０は、未使用領域管理情報から確保領域の登録を解除する。ディスク装置１５０の未使用領域容量が指示された分に満たない場合、未使用領域全体を確保し、別のディスク装置１５０を検索し、条件に合った未使用領域を確保する（ステップ３－３）。

【０１１０】

記憶装置１１０は、ステップ３－３の処理を、指示分の容量のデータ領域を確保するまで繰り返す(ステップ３－４)。

【０１１１】

記憶装置１１０は、該記憶装置が有する論理物理変換テーブルへ、確保された領域を登録することで、確保された領域で構成される論理ディスク装置を定義する(ステップ３－５)。

【０１１２】

記憶装置１１０は、定義した論理ディスク装置の情報を使用者等に報告する(ステップ３－６)。

【０１１３】

記憶装置１１０への未使用領域確保の指示は、第1の実施形態と同じく専用のコマンドを用いてもよいし、第２の実施形態と同じくコマンドボリュームへの命令書き込みでも構わない。保守用に記憶装置１１０へ接続するサービスプロセッサを設け、サービスプロセッサからコマンドを発行する構成も考えられる。

【０１１４】

ステップ１－１からステップ３－４までの一連の処理をスクリプト化することも可能である。その場合、ユーザまたは保守員は、転送先データ領域の選定条件をより詳細に指定することで、転送先データ領域を選定し、データ転送を自動的に実行することが可能となる。転送先データ領域の条件としては、記憶装置１１０の記憶領域での連続性、格納されるディスク装置１５０の物理容量、ヘッド位置付け時間やデータ転送速度などのアクセス性能が考えられる。ＲＡＩＤ装置の場合には、ＲＡＩＤレベルなどの物理構成条件なども転送先データ領域の条件として考えられる。特定のＬＶと物理構成、すなわちディスク装置１５０やディスク装置１５０が接続される内部パスおよび記憶制御プロセッサ１１１を共有しないという条件も考えられる。

【０１１５】

第３実施形態の変形として、転送先領域を指定せずに、転送元ＬＶと転送先領域の選定条件のみをユーザが指定することで、データ転送を指示する方式も考えられる。この場合、記憶装置１１０は、転送先領域を選定条件に従って選定し、新規に生成した論理ディスク装置へデータの転送を行う。記憶装置１１０は、データの転送が完了したら、転送の完了及び転送先として選定した領域情報をホスト１００へ報告する。報告を受けたホスト１００は、報告された転送先論理ディスク装置に対して、ステップ２－１、２－２及び２－４の処理によりＬＶの移動を完了する。このときステップ２－２で、論理ディスク装置内のデータは有効なままＰＶが定義される必要がある。

【０１１６】

なお、本発明は上記の実施形態に限定されず、その要旨の範囲内で数々の変形が可能である。

【０１１７】

各実施形態では、ＰＶ、ホスト１００に提供される論理ディスク装置は、実際のディスク装置１５０と一対一対応であるとしているが、ＰＶは、記憶装置１１０内でＲＡＩＤ５レベルなどのＲＡＩＤで構成されていてもよい。その場合、ホスト１００は、記憶装置１１０が提供する論理ディスク装置へＩ／Ｏを発行する。記憶制御プロセッサ１１１は、論理ディスク装置へのＩ／Ｏを、論理物理変換によりディスク装置１５０へのＩ／Ｏに変換する。

【０１１８】

各実施形態では、データの再配置の例として、ＬＶＭの管理するＬＶの再配置を採用しているが、ＬＶが未割り当てなＰＥを連続化（ガベージコレクション）する処理や、ＤＢＭＳの管理するＤＢ表の再配置の処理など、他のデータ再配置にも本発明を適用することが可能である。

【０１１９】

また、ホスト１００と記憶装置１１０間でのデータ転送処理要求方法について、第１の実施形態の専用コマンドを使用する方法と、第２の実施形態のコマンドボリューム９００を利用する方法は、両実施形態で入れ替えても実現可能である。

【０１２０】

各実施形態では、転送先ＰＶが１台の場合を想定しているが、複数台としても構わない。その場合には、複数台のＰＶのそれぞれについて、転送元データをどのように分散させるかを指定する必要がある。複数のＰＶへの分散の方法としては、各ＰＶに均等に分配する場合、ＰＶの指定順に空き容量の許す限り詰め込んでいく場合が考えられる。また、均等に分配する場合には、さらに、各ＰＶ内でデータを連続させる場合、特定サイズでデータが分割され、ＲＡＩＤにおけるストライピングの如く分割されたデータを各ＰＶに順に格納する場合などが考えられる。

【０１２１】

ユーザ又は保守員があらかじめ転送先として連続領域をＰＥに割り当ててから、データ再配置処理１３１にパラメタとして引き渡す、あるいは、データ再配置

処理内で前述した空きＰＥのガベージコレクションを行っても良い。

【０１２２】

各実施形態では、転送先データ領域へのアクセスは発生しないことを前提としている。つまり、記憶装置１１０での転送先データ領域に対するアクセス抑止の考慮はなく、転送先データ領域へのアクセスがあると、そのままアクセス対象となる領域に対してデータ参照/更新が行われる。しかし、ホスト１００でのアクセス抑止の保証がないケースにそなえて、記憶装置１１０で、データ転送領域の転送先データ領域として登録されているデータ領域に対するI/0を拒否する構成も考えられる。逆に、データ転送処理の完了を待たずに、ホスト１００が、ＬＶ－ＰＶ対応情報１４５を再配置後の状態に更新し、転送対象のＬＶへのアクセスを、転送先ＰＶで受け付ける方法をとってもよい。この場合、記憶装置１１０でデータ転送領域のペアを造り、同期化のためのデータコピーを行うのは第２の実施形態と同じである。ただし、転送先領域へのリード要求に対しては転送元領域のデータに参照し、転送先領域へのライト要求に対しては、転送元領域にもデータの反映を行う必要がある。

【０１２３】

本実施形態によれば、通常よりも短い時間、ＬＶをオフラインとするだけでＬＶの再配置処理をおこなうことができ、システムの可用性を向上することができる。

【０１２４】

【発明の効果】

本発明の計算機システムによれば、記憶装置に格納したデータを別の領域に移動する際に、データ転送処理を記憶装置内で行うことで、ホストおよびチャネルの負荷を削減することが可能となる。

【０１２５】

また、本発明の計算機システムによれば、データ再配置におけるデータ転送中に当該データに対するアクセスを受けつけることが可能となる。この結果、データ再配置における対象データをアクセスする業務の停止時間を短縮することが可能となる。

【図面の簡単な説明】

【図１】

本発明の第１の実施形態が対象とする計算機システムのブロック図である。

【図２】

本発明が想定しているデータ再配置処理の概要図である。

【図３】

本発明におけるＬＶ管理情報の構成図である。

【図４】

本発明におけるＰＶ管理情報の構成図である。

【図５】

本発明におけるデータ転送領域情報の構成図である。

【図６】

本発明の第１の実施形態におけるデータ再配置処理のフロー図である。

【図７】

本発明の第１の実施形態におけるコマンド処理のフロー図である。

【図８】

本発明の第１の実施形態におけるコピー処理のフロー図である。

【図９】

本発明の第２の実施形態における計算機システムの構成図である。

【図１０】

本発明の第２の実施形態におけるデータ再配置処理のフロー図である。

【図１１】

本発明の第２の実施形態におけるコマンド処理のフロー図である。

【図１２】

本発明の第２の実施形態におけるコピー処理のフロー図である。

【符号の説明】

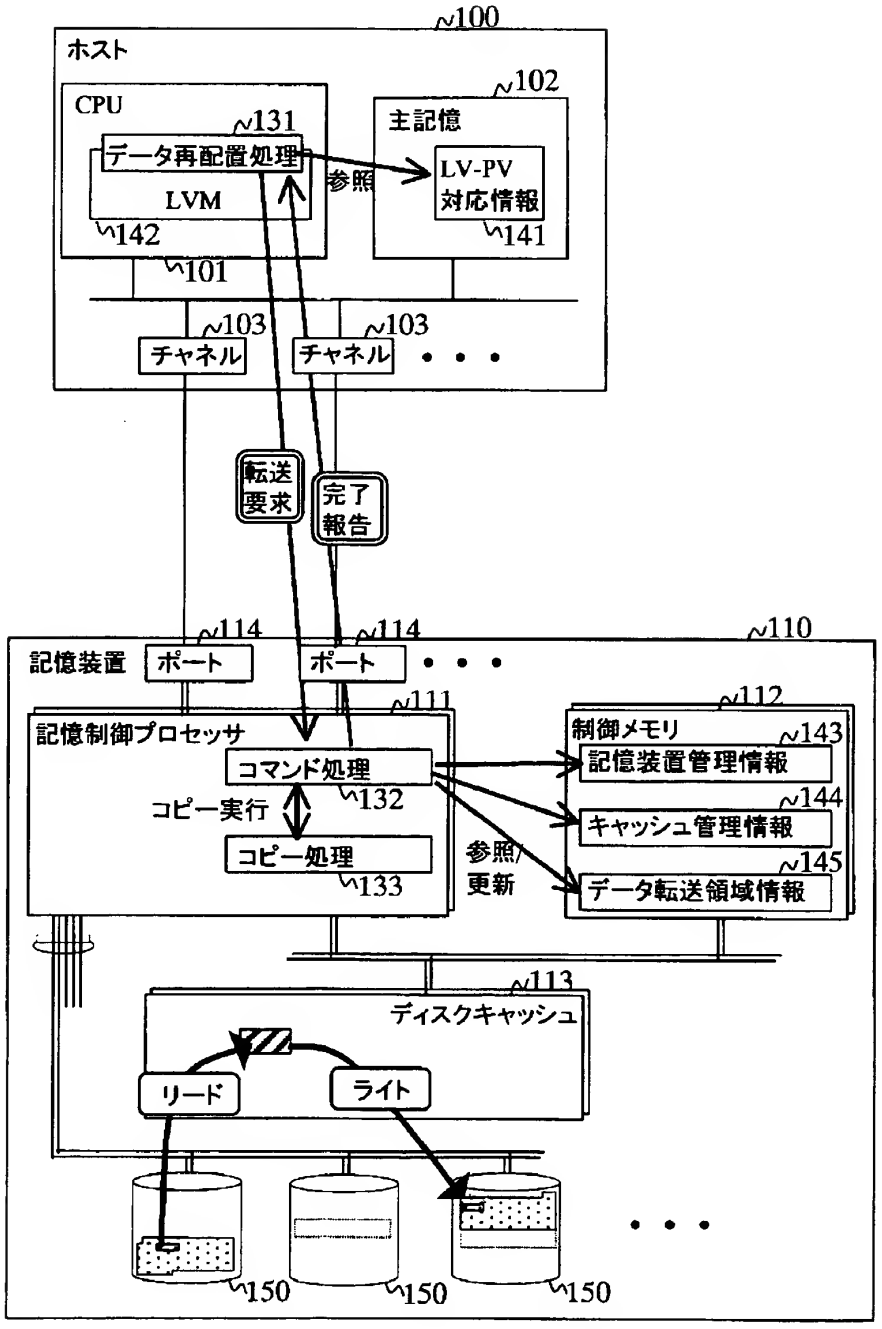
１００…ホスト、１０１…ＣＰＵ、１０２…主記憶、１０３…チャネル、１１０…記憶装置、１１１…記憶制御プロセッサ、１１２…制御メモリ、１１３…ディスクキャッシュ、１４１…ＬＶ－ＰＶ対応情報、１４３…記憶制御管理情報、１

４４…キャッシュ管理情報、１４５…データ転送領域情報、１５０…ディスク装置、９００…コマンドボリューム

【書類名】　図面

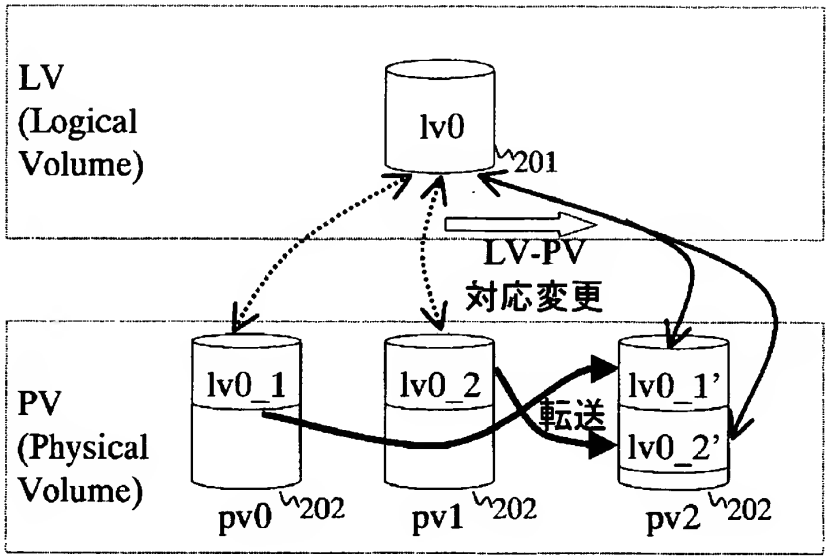
【図１】

図1


100
ホスト
CPU
131
データ再配置処理
LVM
142
参照
102
主記憶
LV-PV
対応情報
141
101
103
チャネル
103
チャネル
転送
要求
完了
報告
114
ポート
114
ポート
110
記憶装置
111
記憶制御プロセッサ
コマンド処理
132
コピー実行
コピー処理
133
参照/
更新
112
制御メモリ
143
記憶装置管理情報
144
キャッシュ管理情報
145
データ転送領域情報
113
ディスクキャッシュ
リード
ライト
150
150
150

【図２】

図2


LV
(Logical
Volume)
lv0
201
LV-PV
対応変更
PV
(Physical
Volume)
lv0_1
lv0_2
転送
lv0_1'
lv0_2'
pv0
202
pv1
202
pv2
202

【図３】

図3

LV管理情報300

PVリスト	～301
LE数	～302
LEサイズ	～303
LE-PE対応情報	～310

LE-PE対応情報310

～311 LE番号	～312 PV名	～313 PE番号
0	/dev/dsk/c0t0d0	500
1	/dev/dsk/c0t0d0	501
2	/dev/dsk/c0t0d0	502
3	/dev/dsk/c0t0d0	503
	• • •	
499	/dev/dsk/c0t0d0	999
500	/dev/dsk/c1t0d0	100
501	/dev/dsk/c1t0d0	101
502	/dev/dsk/c1t0d0	102
	• • •	

【図４】

図4

PV管理情報400

LVリスト	～401
PE数	～402
PEサイズ	～403
PE-LE対応情報	～410

PE-LE対応情報410

PE番号 ～411	LV名 ～412	LE番号 ～413
0	/dev/vg0/lv_other	0
1	/dev/vg0/lv_other	1
2	/dev/vg0/lv_other	2
3	/dev/vg0/lv_other	3
	• • •	
499	/dev/vg0/lv_other	499
500	/dev/vg0/lv0	0
501	/dev/vg0/lv0	1
502	/dev/vg0/lv0	2
	• • •	

【図５】

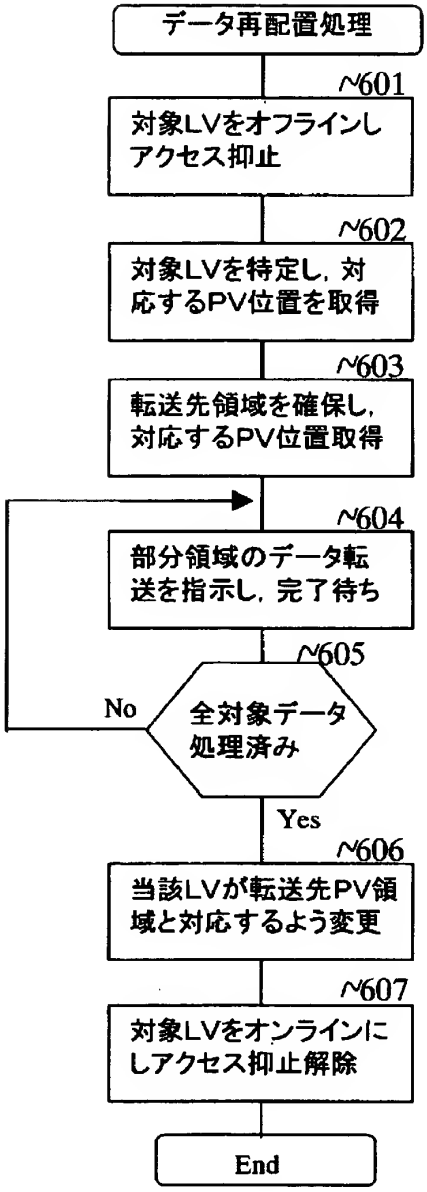
図5

データ転送領域情報145

転送元範囲情報	~501
転送先範囲情報	~502
進捗ポインタ	~503
同期状態	~504
差分ビットマップ	~505

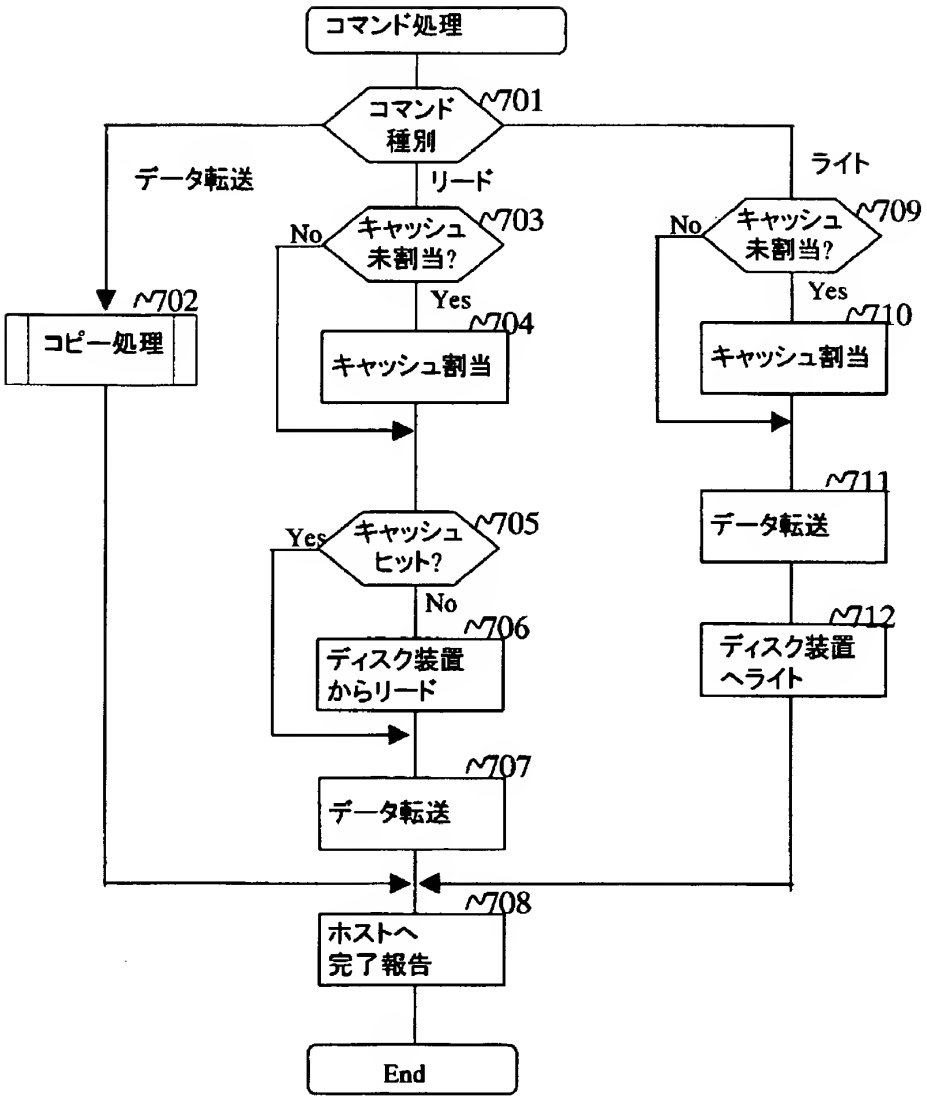
【図６】

図6


データ再配置処理
601
対象LVをオフラインし
アクセス抑止
602
対象LVを特定し，対
応するPV位置を取得
603
転送先領域を確保し，
対応するPV位置取得
604
部分領域のデータ転
送を指示し，完了待ち
605
No
全対象データ
処理済み
Yes
606
当該LVが転送先PV領
域と対応するよう変更
607
対象LVをオンラインに
しアクセス抑止解除
End

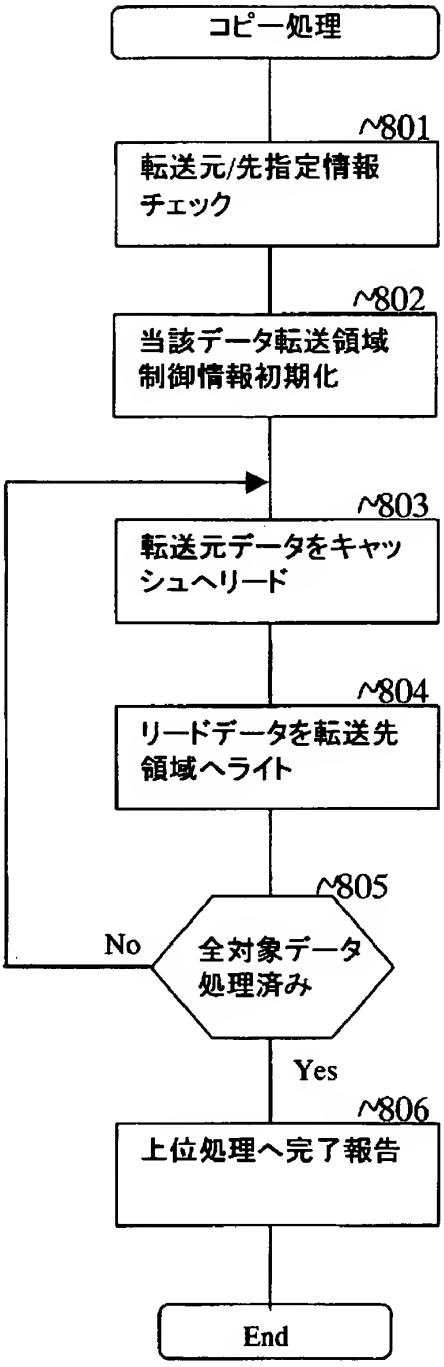
【図７】

図7


コマンド処理
コマンド種別
701
データ転送
リード
ライト
No
キャッシュ未割当?
703
Yes
キャッシュ割当
704
コピー処理
702
No
キャッシュ未割当?
709
Yes
キャッシュ割当
710
Yes
キャッシュヒット?
705
No
ディスク装置からリード
706
データ転送
707
データ転送
711
ディスク装置へライト
712
ホストへ完了報告
708
End

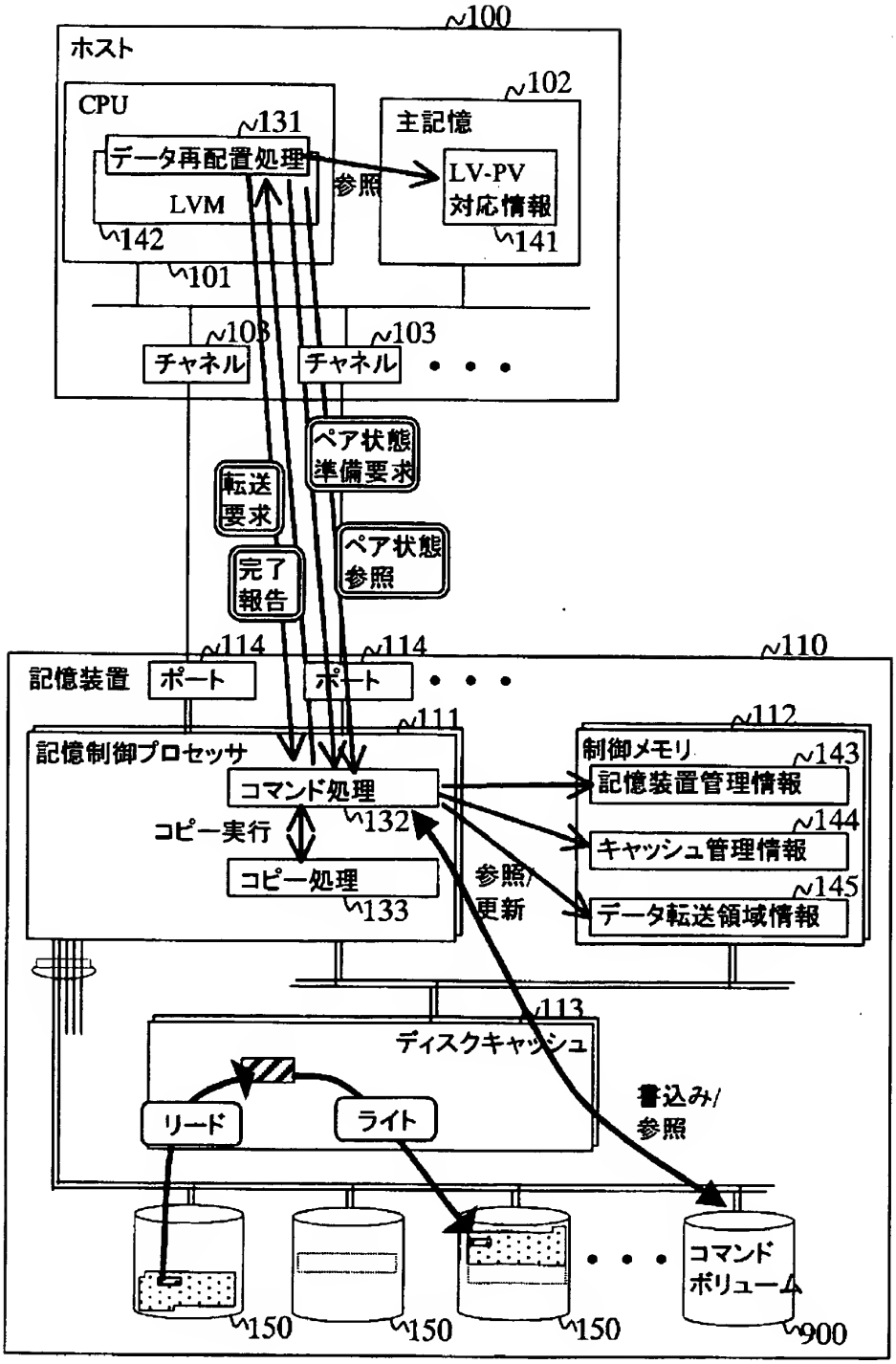
【図８】

図8


コピー処理
801
転送元/先指定情報
チェック
802
当該データ転送領域
制御情報初期化
803
転送元データをキャッ
シュへリード
804
リードデータを転送先
領域へライト
805
No
全対象データ
処理済み
Yes
806
上位処理へ完了報告
End

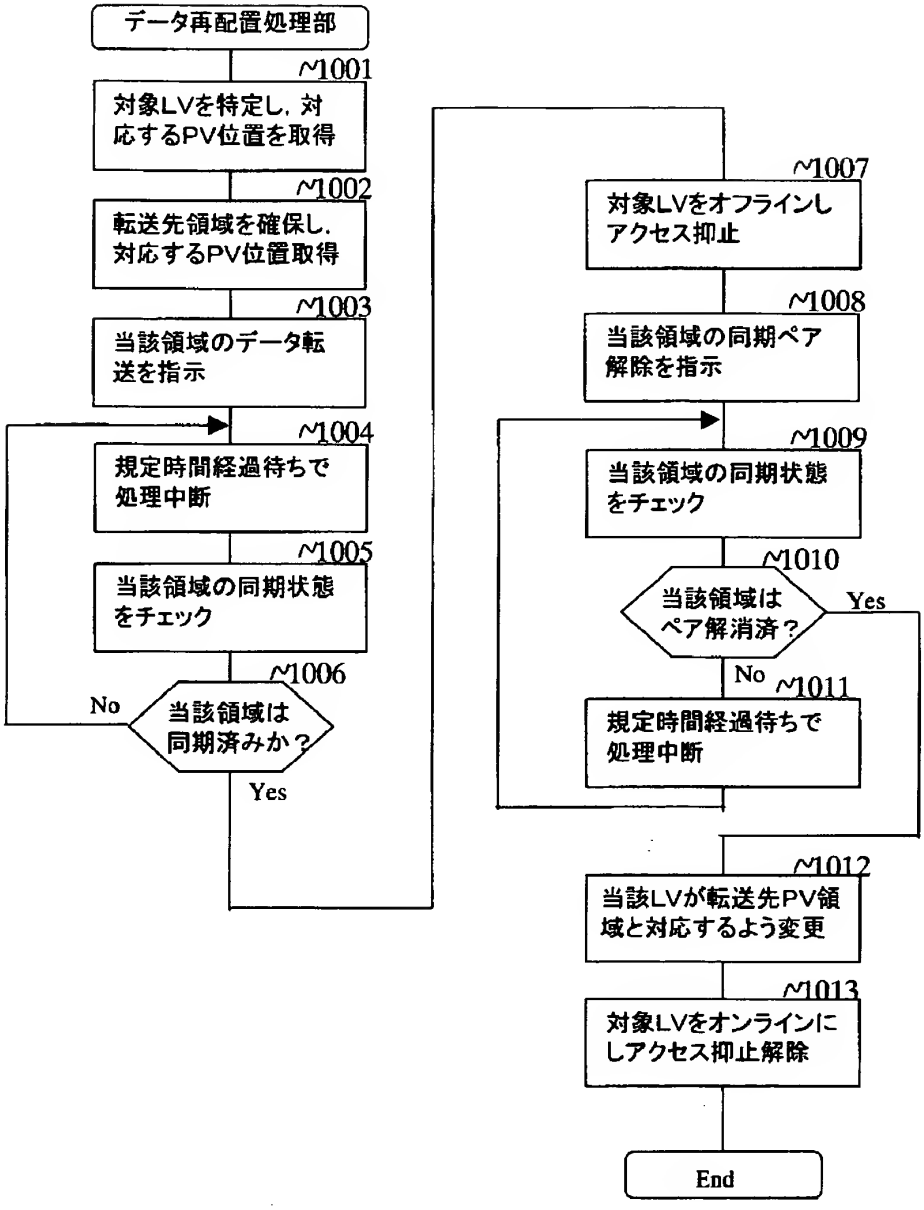
【図９】

図9


100
ホスト
CPU
131
データ再配置処理
LVM
142
参照
102
主記憶
LV-PV
対応情報
141
101
108
チャネル
103
チャネル
ペア状態
準備要求
転送
要求
ペア状態
参照
完了
報告
114
ポート
114
ポート
110
記憶装置
111
記憶制御プロセッサ
コマンド処理
132
コピー実行
コピー処理
133
112
制御メモリ
143
記憶装置管理情報
144
キャッシュ管理情報
145
データ転送領域情報
参照/
更新
113
ディスクキャッシュ
リード
ライト
書込み/
参照
150
150
150
コマンド
ボリューム
900

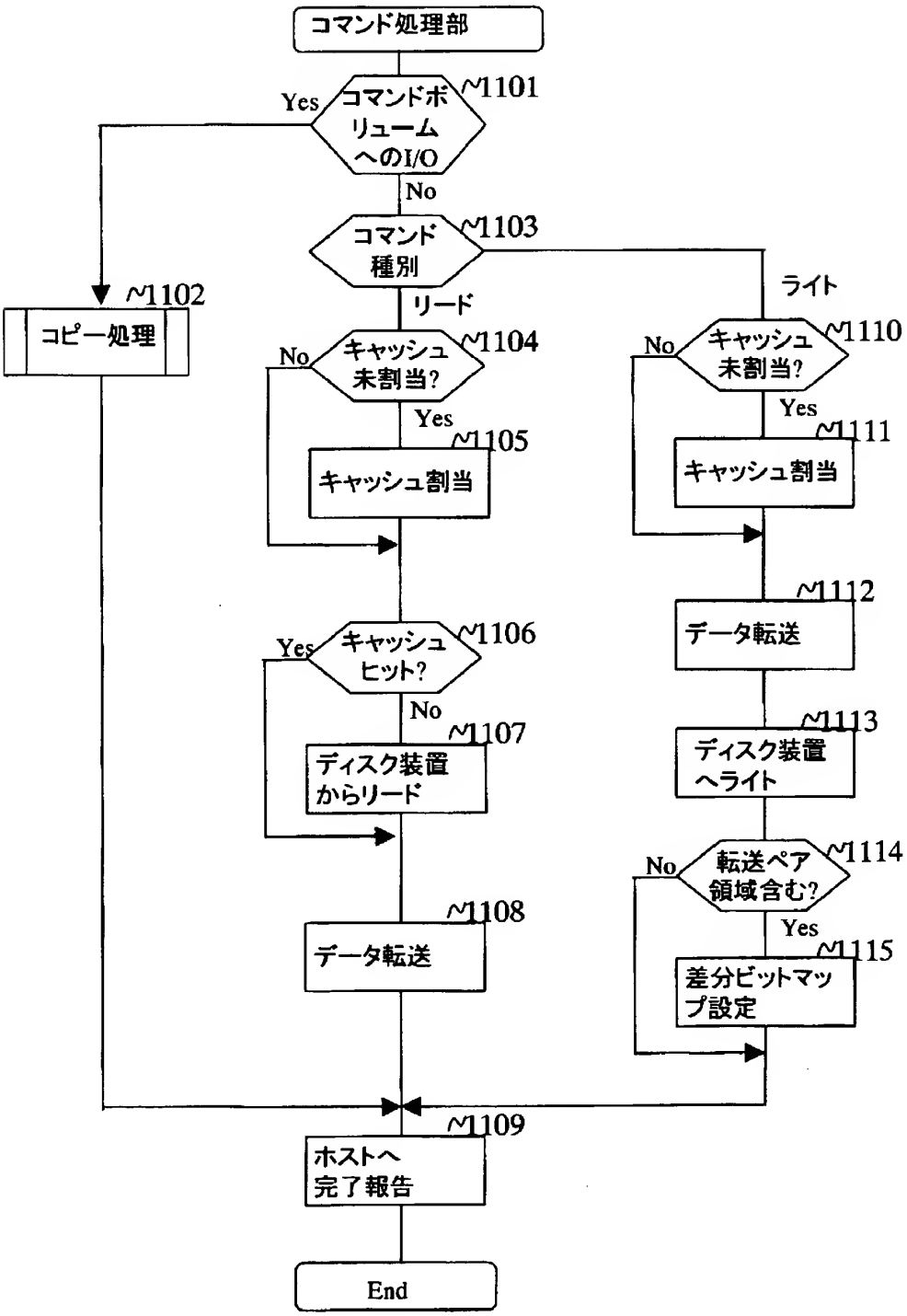
【図１０】

図10


データ再配置処理部
1001
対象LVを特定し，対応するPV位置を取得
1002
転送先領域を確保し，対応するPV位置取得
1003
当該領域のデータ転送を指示
1004
規定時間経過待ちで処理中断
1005
当該領域の同期状態をチェック
1006
当該領域は同期済みか？
No
Yes
1007
対象LVをオフラインしアクセス抑止
1008
当該領域の同期ペア解除を指示
1009
当該領域の同期状態をチェック
1010
当該領域はペア解消済？
Yes
No
1011
規定時間経過待ちで処理中断
1012
当該LVが転送先PV領域と対応するよう変更
1013
対象LVをオンラインにしアクセス抑止解除
End

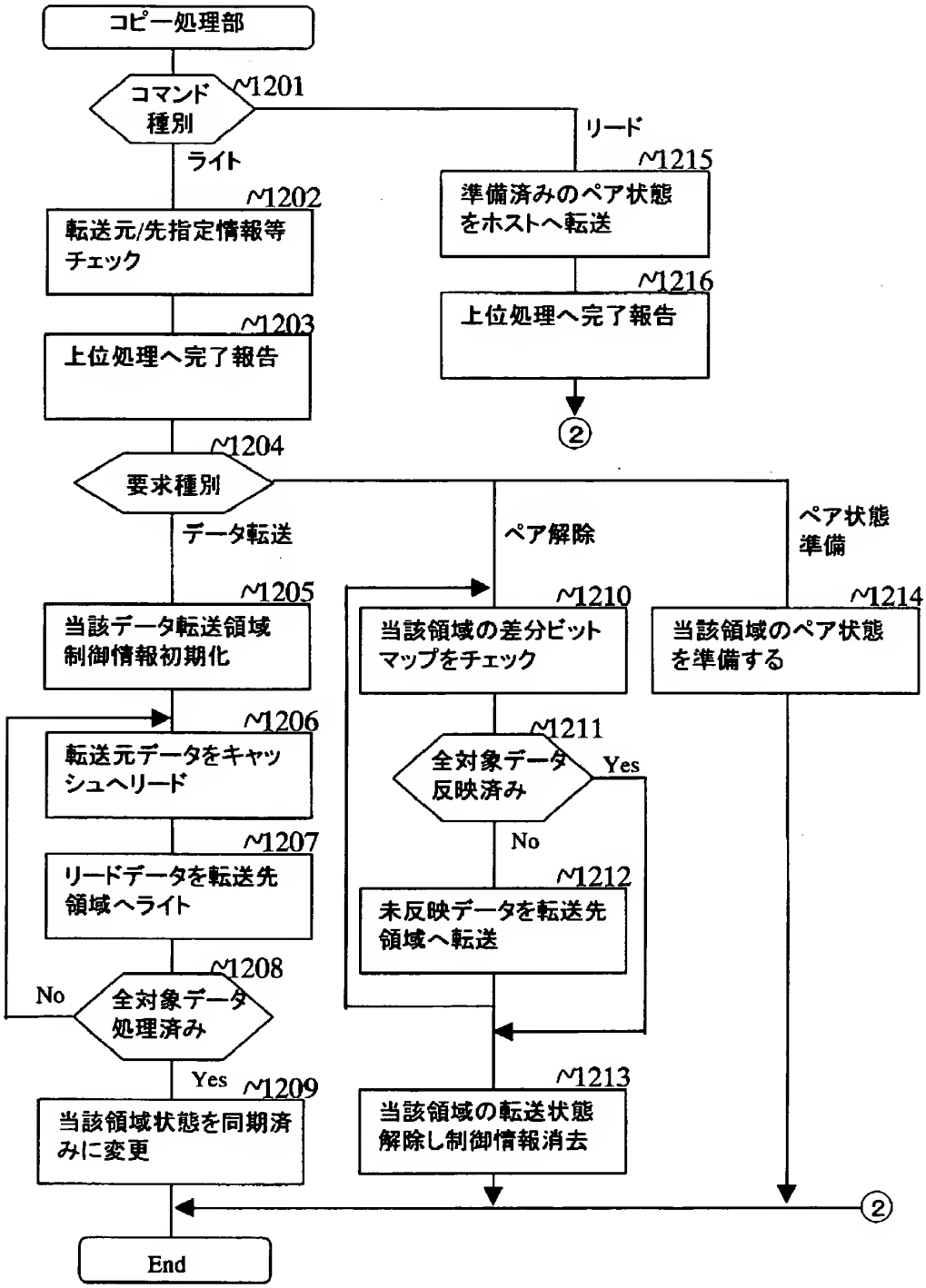
【図１１】

図11


コマンド処理部
Yes
コマンドボリュームへのI/O
1101
No
コマンド種別
1103
ライト
リード
1102
コピー処理
No
キャッシュ未割当?
1104
Yes
キャッシュ割当
1105
No
キャッシュ未割当?
1110
Yes
キャッシュ割当
1111
Yes
キャッシュヒット?
1106
No
ディスク装置からリード
1107
データ転送
1112
ディスク装置へライト
1113
No
転送ペア領域含む?
1114
Yes
差分ビットマップ設定
1115
ホストへ完了報告
1109
End
1108

【図１２】

図12


コピー処理部
コマンド種別
1201
リード
ライト
1215
準備済みのペア状態をホストへ転送
1202
転送元/先指定情報等チェック
1216
上位処理へ完了報告
1203
上位処理へ完了報告
②
1204
要求種別
データ転送
ペア解除
ペア状態準備
1205
当該データ転送領域制御情報初期化
1210
当該領域の差分ビットマップをチェック
1214
当該領域のペア状態を準備する
1206
転送元データをキャッシュへリード
1211
全対象データ反映済み
Yes
No
1207
リードデータを転送先領域へライト
1212
未反映データを転送先領域へ転送
1208
No
全対象データ処理済み
Yes
1209
当該領域状態を同期済みに変更
1213
当該領域の転送状態解除し制御情報消去
②
End

【書類名】　要約書

【要約】

【課題】

ストレージに格納したデータを別の領域に再配置する処理において、領域間のデータ転送をホストで行うため、ホストおよびチャネルに負荷がかかる。

【解決手段】

ホストはデータ転送元/先の領域情報をパラメタとしてストレージへデータ転送を依頼する。ストレージは内部で転送元のディスク装置から転送先のディスク装置へデータ転送を行う。対象データの転送が全て完了したら、ホストはそのデータの格納場所を転送先領域に変更して登録する。

【選択図】　図１

特願２００１－０５３４７２

出願人履歴情報

識別番号 [０００００５１０８]

１．変更年月日 １９９０年 ８月３１日
[変更理由] 新規登録
住所 東京都千代田区神田駿河台４丁目６番地
氏名 株式会社日立製作所